Priors and Hierarchical Bayesian Modeling

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

January 19, 2019

References

- C P Robert, The Bayesian Choice: From Decision-Theoretic Motivations to Compulational Implementation, Springer-Verlag, NY, 2001 (online resource)
- A Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis, Chapman and Hall CRC Press, 2nd Edition, 2003.
- J M Marin and C P Robert, The Bayesian Core, Spring Verlag, 2007 (online resource)
- D. Sivia and J Skilling, Data Analysis: A Bayesian Tutorial, Oxford University Press, 2006.
- Bayesian Statistics for Engineering, Online Course at Georgia Tech, B. Vidakovic.
- Kevin Murphy, Machine Learning, A probabilistic Perspective, Chapter 5.

Contents

Selection of Prior Distribution

- Once the prior distribution is selected, Bayesian inference can be performed almost mechanically.

- A critical point of Bayesian statistics is the choice of the prior.

- Seldom there is enough “subjective information” to lead to an `exact’ determination of the prior distribution.

- Selection of prior includes subjectivity

  - ✓ Subjectivity does not imply being unscientific – one can use scientific information to guide the specification of priors.

  - ✓ We will review some of the work on uninformative and robust priors.

Informative Priors

- The prior is a tool summarizing the available information on a phenomenon of interest, as well as the uncertainty related with this information.

- Informative priors convey specific and definite information about parameters θ associated with the random phenomenon.

- Pre-existing evidence which has already been taken into account is part of the informative priors. This information can be based on historical data, insight or personal beliefs.

- Typical subgroups of informative priors

  - conjugate, non-conjugate

  - exponential families

  - maximum entropy priors

Conjugate Priors

- Consider a class of probability distributions P. For every prior $\pi(\theta) \epsilon P$, if the posterior distribution $\pi(\theta|x)$ belongs to P and the likelihood $f(x|\theta)$ to a family F, then the ***P* class is conjugate for *F***.

- Conjugate priors are analytically tractable. Finding the posterior reduces to an updating of the corresponding parameters of the prior.

- Consider a coin flipping example:

  - Let θ the probability that the coin will draw heads
  - Prior $\theta \sim \mathscr{Be}(a,b)$
  - Data: the coin flipped n times with n_H of those were heads (binomial)
  - Posterior:

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_0^1 f(x \mid \theta)\pi(\theta)d\theta} = \frac{\theta^{a+n_H-1}(1-\theta)^{b+n-n_H-1}}{beta(a+n_H, b+n-n_T)} = \mathscr{Be}(a+n_H, b+n-n_H)$$

- The role of conjugate priors is generally to provide a first approximation to the adequate prior distribution which should be followed by a robustness analysis.

Exponential Family

- Conjugate prior distributions are usually associated with the **Exponential Family**, a class of probability distributions sharing a certain form as specified below.

- Suppose $\boldsymbol{x}$ are observations from the Exponential Family

$$f(\mathbf{x}\,|\,\theta) = C(\theta)h(\mathbf{x})\exp\left\{R(\theta)\cdot T(\mathbf{x})\right\}$$

We call this an exponential family. $T(\boldsymbol{x})$ are sufficient statistics.

- When $\Theta \subset \mathbb{R}^k, X \subset \mathbb{R}^k$ *and*

$$f(\mathbf{x}\,|\,\theta) = h(\mathbf{x})\exp\left\{\theta\cdot\mathbf{x} - \psi(\theta)\right\}$$

the family is called natural family of dimension k.

Exponential Family: Example

- Consider the likelihood function

$$f(x|\theta)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\theta)^2}{2}\right\}$$

- This is a normal distribution (unknown mean, unit variance). For this case note that:

$$f(\mathbf{x}|\theta)=h(\mathbf{x})\exp\left\{\theta\cdot\mathbf{x}-\psi(\theta)\right\} \qquad R(\theta)=\theta\ ;\ T(x)=x\,;\psi(\theta)=\frac{\theta^2}{2}\ ;\ h(x)=\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$$

- Consider the normal distribution (unknown mean, unknown variance)

$$f(x|\theta)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

define $\theta=(\mu,\sigma)$, we can then see that

$$f(\mathbf{x}|\theta)=C(\theta)h(\mathbf{x})\exp\left\{R(\theta)\cdot T(\mathbf{x})\right\} \qquad R(\theta)=(\frac{\mu}{\sigma^2},\frac{1}{\sigma^2})^T\ ;\ T(x)=(x,-\frac{x^2}{2})^T\,;C(\theta)=\frac{1}{\sigma}e^{-\frac{\mu^2}{2\sigma^2}};$$

$$f(\mathbf{x}|\theta)=h(\mathbf{x})\exp\left\{R(\theta)\cdot T(\mathbf{x})-\psi(\theta)\right\} \qquad \psi(\theta)=\frac{\mu^2}{2\sigma^2}-\log\frac{1}{\sigma};h(x)=\frac{1}{\sqrt{2\pi}}.$$

Exponential Family

- Conjugate distributions for exponential families

  - Likelihood

$$f(\boldsymbol{x}\,|\,\theta) = h(\boldsymbol{x})\exp\left\{R(\theta)\cdot T(\boldsymbol{x}) - \psi(\theta)\right\}$$

  - Conjugate Prior

$$\pi(\theta\,|\underbrace{\mu,\lambda}_{\textit{Hyper Parameters}}) \propto \exp\left\{R(\theta)\cdot\mu - \lambda\psi(\theta)\right\},\ \lambda > 0$$

  - Posterior

$$\pi(\theta\,|\,\boldsymbol{x}) \propto \exp\left\{R(\theta)\cdot\left[\mu + T(\boldsymbol{x})\right] - (\lambda+1)\psi(\theta)\right\}$$

$$i.e.\quad \pi(\theta\,|\,\boldsymbol{x}) = \pi(\theta\,|\,\mu + T(\boldsymbol{x}), \lambda+1)$$

Exponential Family: Example

- Normal distribution (unknown mean, known variance)

$$Likelihood: x_1 \mid \theta \sim \mathcal{N}(\theta, \sigma^2), \sigma^2 = known, x_1 \in \mathbb{R}$$

$$f(x_1 \mid \theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_1-\theta)^2}{2\sigma^2}}$$

  - Conjugate prior

$$\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

  - Posterior

$$\theta \mid x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2),\ \sigma_1^{-2} = \sigma_0^{-2} + \sigma^{-2},\ \mu_1 = \underbrace{\frac{\sigma_0^{-2}\mu_0 + \sigma^{-2}x_1}{\sigma_0^{-2} + \sigma^{-2}}}_{\text{weighted average of the observation } x_1 \text{ and the prior mean}}$$

  - Posterior predictive:

$$\pi(x \mid x_1) = \int \pi(x \mid \theta)\pi(\theta \mid x_1) d\theta \sim \int e^{-\frac{(x-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}} d\theta \sim \mathcal{N}(\mu_1, \sigma^2 + \sigma_1^2)$$

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004

Gaussian With Multiple Observations - Unknown Mean

- Assume we have observations $X_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$ and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- The posterior is then:

$$\mu \mid x_1, x_2, \ldots, x_n \sim \mathcal{N}(\mu_n, \sigma_n^2),$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \Rightarrow \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\sigma^2}{n + \sigma^2 / \sigma_0^2}$$

$$\mu_n = \sigma_n^2 \left(\frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \sigma_n^2 \left(\frac{\sum_{i=1}^n x_i + \mu_0 \left(\sigma^2 / \sigma_0^2 \right)}{\sigma^2} \right)$$

- One can think of the prior as n_0 virtual observations with $n_0 = \frac{\sigma^2}{\sigma_0^2}$ and

$$\sigma_n^2 = \frac{\sigma^2}{n + n_0}, \quad \mu_n = \frac{\sum_{i=1}^n x_i + n_0 \mu_0}{n + n_0}$$

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004

Standard Exponential Family of Distributions

f(x|θ)	π(θ)	π(θ|x)
Normal $N(\theta,\sigma^2)$	Normal $\mathcal{N}(\mu,\tau^2)$	$\mathcal{N}\left(\rho\left(\sigma^2\mu+\tau^2x\right),\rho\sigma^2\tau^2\right)$ $\rho^{-1}=\sigma^2+\tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{P}(\theta)\mathcal{G}\left(\alpha,\beta\right)$	$\mathcal{G}(\alpha+x,\beta+1)$
Gamma $\mathcal{G}(v,\theta)$	Gamma $\mathcal{G}(\alpha,\beta)$	$\mathcal{G}(\alpha+v,\beta+x)$
Binomial $\mathcal{B}(\mathrm{n},\theta)$	Beta $\mathcal{B}\mathrm{e}(\alpha,\beta)$	$\mathcal{B}\mathrm{e}(\alpha+x,\beta+n-x)$
Negative Binomial $\mathcal{N}\mathrm{eg}(\mathrm{m},\theta)$	Beta $\mathcal{B}\mathrm{e}(\alpha,\beta)$	$\mathcal{B}\mathrm{e}(\alpha+m,\beta+x)$
Multinomial $\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	Dirichlet $\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal $\mathcal{N}(\mu,1/\theta)$	Gamma $\mathcal{G}\mathrm{a}(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

Mixture of Conjugate Priors

- Robust priors are useful, but can be computationally expensive to use.

- Conjugate priors simplify the computation, but are often not robust, and not flexible enough to encode our prior knowledge.

- A mixture of conjugate priors is also conjugate and can approximate any kind of prior. Thus such priors provide a good compromise between computational convenience and flexibility.

- Example: to model coin tosses, we can take a prior which is a mixture of two beta distributions to model coin tosses.

$$p(\theta) = 0.5\mathcal{B}eta(\theta \mid 20, 20) + 0.5\mathcal{B}eta(\theta \mid 30, 10)$$

- If θ comes from the first distribution, the coin is fair, but if it comes from the second, it is biased towards heads.

Mixture of Conjugate Priors

- If we have a prior distribution which is a mixture of conjugate distributions to a given likelihood, then the posterior is in closed form and is a mixture of conjugate distributions, i.e. with

$$\pi(\theta)=\sum_{i=1}^{K} w_i \pi_i(\theta) \equiv \sum_{i=1}^{K} P(Z=i)\pi(\theta \mid Z=i)$$

we obtain $\pi(\theta \mid \mathscr{D})=\dfrac{\sum_{i=1}^{K} w_i \pi_i(\theta) f(\mathscr{D} \mid \theta)}{\underbrace{\sum_{i=1}^{K} w_i \int \pi_i(\theta) f(\mathscr{D} \mid \theta) d\theta}_{A}}=\sum_{i=1}^{K} \dfrac{w_i}{A} \pi_i(\theta) f(\mathscr{D} \mid \theta)$ or

$$\pi(\theta \mid \mathscr{D})=\sum_{i=1}^{K} w_i' \frac{\pi_i(\theta) f(\mathscr{D} \mid \theta)}{\int \pi_i(\theta) f(\mathscr{D} \mid \theta) d\theta}=\sum_{i=1}^{K} w_i' \pi_i(\theta \mid \mathscr{D}) \equiv \sum_{i=1}^{K} P(Z=i \mid \mathscr{D}) \pi(\theta \mid \mathscr{D}, Z=i)$$

where:

$$p(Z=i \mid \mathscr{D})=\frac{p(Z=i) p(\mathscr{D} \mid Z=i)}{\sum_k p(Z=k) p(\mathscr{D} \mid Z=k)}=\frac{w_i \int \pi_i(\theta) f(x \mid \theta) d\theta}{\sum_{k=1}^{K} w_k \int \pi_k(\theta) f(x \mid \theta) d\theta}=w_i', \quad \sum_{i=1}^{K} w_i'=1.$$

- One can approximate arbitrary closely any prior distribution by a mixture of conjugate distributions ([Brown, 1986](#))

Mixture of Conjugate Priors

- As an example, suppose we use the mixture prior

$$p(\theta) = 0.5\mathcal{Beta}(\theta \mid a_1, b_1) + 0.5\mathcal{Beta}(\theta \mid a_2, b_2)$$

$$a_1 = b_1 = 20,\ a_2 = b_2 = 10, \textit{ we observe } N_1 \textit{ heads, } N_0 \textit{ tails}$$

- The posterior becomes

$$p(\theta \mid \mathcal{D}) = p(Z=1 \mid \mathcal{D})\mathcal{Beta}(\theta \mid a_1 + N_1, b_1 + N_0) + p(Z=2 \mid \mathcal{D})\mathcal{Beta}(\theta \mid a_2 + N_1, b_2 + N_0)$$

- The posterior mixing weights are given as:

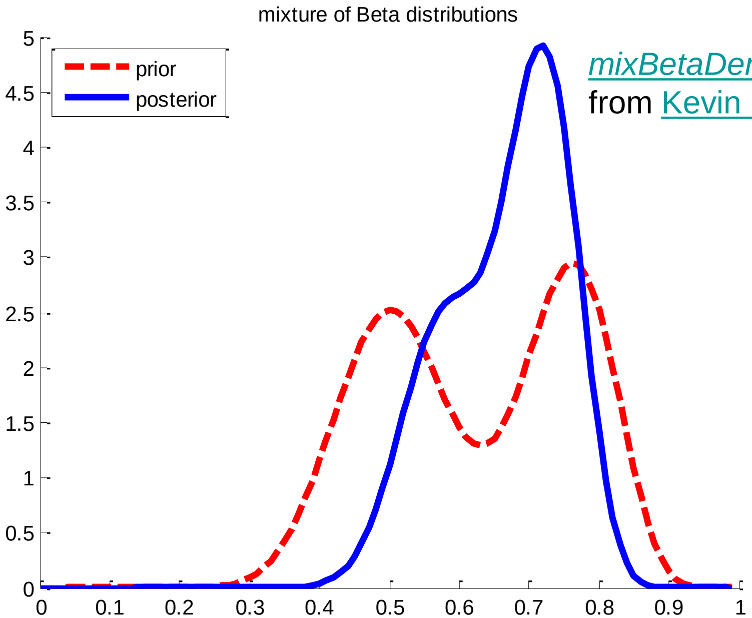
$$p(Z=k \mid \mathcal{D}) = \frac{p(Z=k)\,p(\mathcal{D} \mid Z=k)}{\sum_{k'} p(Z=k')\,p(\mathcal{D} \mid Z=k')} = \frac{p(Z=k)\,p(\mathcal{D} \mid Z=k)}{p(\mathcal{D})}$$

- If $N_1 = 20$ heads and $N_0 = 10$ tails, then, using

$$p(\mathcal{D} \mid Z=1) = \binom{N}{N_1} \frac{B\left(a_1 + N_1, b_1 + N_0\right)}{B\left(a_1, b_1\right)}$$

- The posterior finally becomes

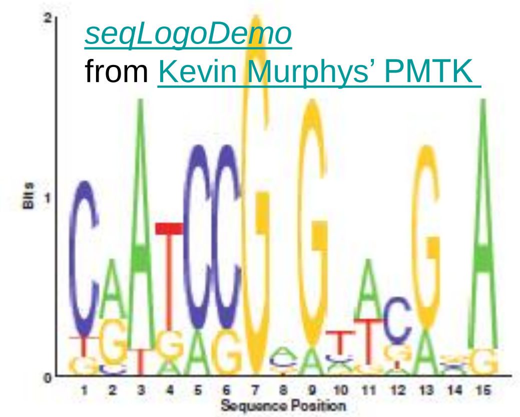
$$p(\theta \mid \mathcal{D}) = 0.346\mathcal{Beta}(\theta \mid 40, 30) + 0.654\mathcal{Beta}(\theta \mid 30, 20)$$



mixBetaDemo from Kevin Murphys' PMTK

Mixture of Conjugate Priors

- Dirichlet-multinomial models are widely used in biosequence analysis. Consider the sequence logo problem.



- Suppose we want to find locations which represent coding regions of the genome. Such locations often have the same letter across all sequences (mostly all A's, or all T's, or all C's, or all G's).

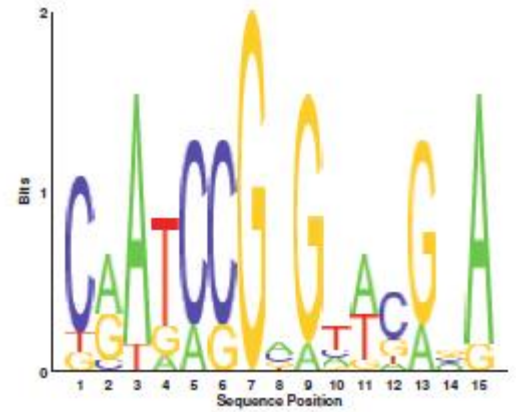
- We believe adjacent locations are conserved together. We let $Z_t = 1$ if location *t* is conserved, and let $Z_t = 0$ otherwise. We add a dependence between adjacent Z_t variables using a Markov chain.

- To define a likelihood model, $p(\boldsymbol{N}_t|Z_t)$, where $\boldsymbol{N}_t$ is the vector of (A, C, G, T) counts for column t. We make this a multinomial distribution with parameter $\boldsymbol{\theta}_t$.

- Since each column has a different distribution, we will want to integrate out $\boldsymbol{\theta}_t$ and thus compute the marginal likelihood $p(\boldsymbol{N}_t|Z_t)$.

Mixture of Conjugate Priors

$$p(\boldsymbol{N}_t \mid Z_t) = \int p(\boldsymbol{N}_t \mid \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t \mid Z_t) d\boldsymbol{\theta}_t$$



- But what prior should we use for $\boldsymbol{\theta}_t$?

- *When $Z_t = 0$ we can use a uniform prior*, $p(\boldsymbol{\theta}|Z_t = 0) = \mathcal{Dir}(1,1,1,1)$, but what should we use if $Z_t = 1$?

- If the column is conserved, $Z_t = 1$, it could be a nearly pure column of A's, C's, G's, or T's. A natural approach is to use a mixture of Dirichlet priors, each tilted towards the appropriate corner of the 4-d simplex,

$$p(\boldsymbol{\theta}|Z_t = 1) = 1/4\,\mathcal{Dir}(\boldsymbol{\theta}|(10,1,1,1)) + \cdots + 1/4\,\mathcal{Dir}(\boldsymbol{\theta}|(1,1,1,10))$$

- Since this is conjugate, we can easily compute $p(\boldsymbol{N}_t|Z_t)$ (Brown et al. 1993)

Summary: Conjugate Priors

- **PROS.**
  - Simple to handle, can be interpreted through imaginary observations.
  - Considered as the least informative ones.
- **CONS.**
  - Not applicable to all likelihood functions.
  - Not flexible, cannot account for constraints e.g. $\theta > 0$.
  - Approximation by mixtures while feasible is very tedious and thus not used in practice.

Noninformative Priors

- The motivation for noninformative priors

  - when prior information about the model is too vague or unreliable, it is usually impossible to justify the choice of prior distributions on a subjective basis.

  - "Objectivity" requirements which force us to provide prior distributions with as little subjective input as possible, in order to base inference on the sampling model alone.

- An intrinsic and acceptable notion of noninformative priors should satisfy invariance under reparametrization.

Noninformative Priors

- Noninformative priors are intended to have as little influence on the posterior as possible i.e. 'letting the data speak for themselves'.

- Assume a distribution $p(x|\lambda)$ governed by a parameter λ, and a prior $p(\lambda) = const$ e.g. if λ is a discrete variable with K states, this simply amounts to setting the prior probability of each state to $1/K$.

- In the case of continuous λ there are two difficulties with this approach. If the domain of λ is unbounded, this prior distribution cannot be correctly normalized (improper prior).

- Improper priors can often be used provided the corresponding posterior distribution is proper.

  - For example, if we put a uniform prior distribution over the mean of a Gaussian, then the posterior distribution for the mean, once we have observed at least one data point, will be proper.

Noninformative Priors

- If we don't have strong beliefs about what θ should be, it is common to use an uninformative prior, and to let the data speak for itself.

- Consider as an example a Bernoulli parameter, $\theta \in [0,1]$.

- An uninformative prior would be the uniform distribution, $\mathcal{Beta}(1,1)$. In this case, the posterior mean and MLE are:

$$\mathbb{E}[\theta \mid \mathcal{D}] = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

$$\bar{\theta} = \frac{N_1}{N_1 + N_0}$$

- *One could argue that the prior wasn't completely uninformative after all.*

$$\mathcal{Beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \mathbb{E}[x] = \frac{\alpha}{\alpha+\beta}$$

Noninformative Priors

- By the above argument, the most non-informative prior is

$$\lim_{c \to 0} \mathcal{Beta}(c,c) = \mathcal{Beta}(0,0)$$

- This prior is a mixture of two equal point masses at 0 and 1.

- It is called *the Haldane prior*.

- Note that the Haldane prior is an improper prior, meaning it does not integrate to 1. However, as long as we see at least one head and at least one tail, the posterior will be proper.

- We will see shortly that the right uninformative prior is:

$$\mathcal{Beta}(1/2, 1/2)$$

$$\mathcal{Beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \mathbb{E}\left[x\right] = \frac{\alpha}{\alpha+\beta}$$

Noninformative Priors

- A second difficulty arises from the transformation behavior of a probability density under a nonlinear change of variables.

- If a function $h(\lambda)$ is constant, and we change variables to $\lambda = \eta^2$, then $h(\eta) = h(\eta^2)$ will also be constant. However, if we choose the density $p_\lambda(\lambda)$ to be constant, then the density of η will be given by

$$p_\eta(\eta) = p_\lambda(\lambda)\left|\frac{d\lambda}{d\eta}\right| = p_\lambda(\eta^2)2\eta \propto \eta$$

  and so the density over η will not be constant.

- This issue does not arise when we use maximum likelihood, because the likelihood function $p(x|\lambda)$ is a simple function of λ and so we are free to use any convenient parameterization.

- If, however, we are to choose a prior distribution that is constant, we must take care to use an appropriate representation for the parameters.

Translation Invariant Prior

- Translation Invariant: Consider a density of the form

$$p(x \mid \mu) = f(x-\mu)$$

  then $f(.)$ is translation invariant and μ is a location parameter.

- Note that if we shift x by a constant to give $\bar{x} = x + c$ then

$$p(\bar{x}|\bar{\mu}) = f(\bar{x} - \bar{\mu}),\ where\ \bar{\mu} = \mu + c$$

- Thus the form of the density remains the same.

- We would like to find a prior that satisfies this translational invariance – a density independent of the origin.

Translation Invariant Prior

- We want a prior that assigns equal probability to the interval $A \leq \mu \leq B$ as to the interval $A - c \leq \mu \leq B - c$.

$$\int_A^B p(\mu)d\mu = \int_{A-c}^{B-c} p(t)dt \underset{t=\mu-c}{=} \int_A^B p(\mu-c)d\mu$$

- A translation invariance requirement is thus that the prior distribution should satisfy:

$$p(\mu) = p(\mu - c) \;\; \text{for every } c \in \mathbb{R} \Rightarrow$$

$$p(\mu) = \text{constant (improper prior)}$$

- This flat prior is improper – but the resulting posterior is proper assuming

$$\int f(x-\theta)d\theta < \infty$$

Having seen $N \geq 1$ data points will satisfy this. One data point is enough to fix the location.

- Example of a location parameter is the mean μ of a Gaussian. The noninformative prior is obtained from the conjugate prior

$$\mathcal{N}(\mu \mid \mu_0, \sigma_0^2) \text{ with } \sigma_0^2 \to \infty.$$

Scale Invariant Prior

- Scale Invariant: If the density is of the form

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

then $f(.)$ is scale invariant and σ is the scale parameter.

- Note that if we change the scale by a constant to give $\bar{x} = cx$ then

$$p(\bar{x}|\bar{\sigma}) = \frac{1}{\bar{\sigma}} f\left(\frac{\bar{x}}{\bar{\sigma}}\right), \; where \; \bar{\sigma} = c\sigma$$

- Thus the form of the density remains the same.

- We would like to find a prior that satisfies this scale invariance – a density independent of the scaling used.

Scale Invariant Prior

- We want a prior that assigns equal probability to the interval $A \le \sigma \le B$ as to the interval $A/c \le \sigma \le B/c$.

$$\int_A^B p(\sigma)d\sigma = \int_{A/c}^{B/c} p(t)dt \underset{t=\frac{\sigma}{c}}{=} \int_A^B p(\frac{\sigma}{c})\frac{1}{c}d\sigma$$

- A translation invariance requirement is thus that the prior distribution should satisfy:

$$p(\sigma) = p(\frac{\sigma}{c})\frac{1}{c} \quad \text{for every } c \in \mathbb{R} \Rightarrow$$

$$p(\sigma) \propto \frac{1}{\sigma} \text{(improper prior)} \Leftrightarrow \mathrm{p}\left(\ln\sigma\right) = const$$

- We can approximate this with a $p(\sigma) = \mathcal{Gamma}(\sigma \mid 0,0)$. *This improper prior leads to a proper posterior if we observe $N \geq 2$ data* (we need at least 2 data points to estimate a variance)

Scale Invariant Prior

- Example of a scale parameter is the std σ of a Gaussian after we account for the location parameter:

$$\mathcal{N}(x|\mu,\sigma^2) \propto \frac{1}{\sigma} e^{-\left(\frac{\tilde{x}}{\sigma}\right)^2}, \tilde{x} = x - \mu$$

- We can express this in terms of the precision $\lambda = 1/\sigma^2$ rather than σ itself.

- A distribution $p(\sigma) \propto 1/\sigma$ corresponds to a distribution over λ of the form $p(\lambda) \propto 1/\lambda$.

- The conjugate prior for λ is $\mathcal{Gamma}(\lambda|a_0,b_0)$. The noninformative prior is obtained from the $\mathcal{Gamma}$ with $a_0 = b_0 = 0$. In this case, the posterior depends only from the data and not from the prior.

$$p(\lambda|\boldsymbol{X},\mu) = \prod_{n=1}^{N} f(x_n|\mu)\boldsymbol{\mathcal{Gamma}}(\lambda|a_0,b_0) \propto \lambda^{N/2+a_0-1}\exp\left(-b_0\lambda - \frac{1}{2}\lambda\sum_{n=1}^{N}(x_n-\mu)^2\right)$$

Jeffrey's Noninformative Priors

- Jeffrey's proposes a more intrinsic approach which avoids the need to take the invariance structure into account.

- Given a likelihood $f(\boldsymbol{x}|\theta)$, Jeffrey's noninformative prior distributions are based on Fisher information, given by

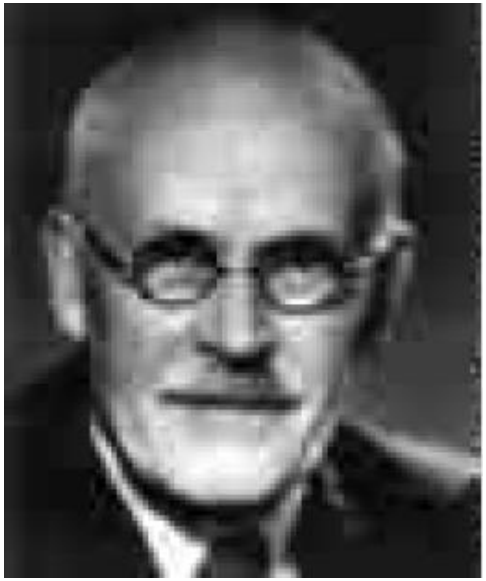
$$I(\theta) = \mathbb{E}_{X|\theta}\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\frac{\partial \log f(X|\theta)^T}{\partial \theta}\right) = -\mathbb{E}_{X|\theta}\left(\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\right)$$

the corresponding prior distribution is

$$\pi(\theta) \propto \left|I(\theta)\right|^{1/2}$$

Determinant of I

Sir Harold Jeffreys
(1891–1989)



Jeffrey's Noninformative Priors

- Jeffreys Invariance Principle:

  - Any rule for defining the prior distribution on θ should lead to an equivalent result when using a transformed parameterization

  - Let $\phi = h(\theta)$ and h be an invertible function with inverse function $\theta = g(\phi)$, then

$$\pi(\phi) = \pi(g(\phi))\left|\frac{dg(\phi)}{d\phi}\right| = \pi(\theta)\left|\frac{d\theta}{d\phi}\right|$$

  - Jeffreys noninformative priors $\pi(\phi) \propto \left|I(\phi)\right|^{1/2}$ satisfy this invariant reparameterization requirement.

$$I(\phi) = -\mathbb{E}_{X|\phi}\left(\frac{\partial^2 \log f(X\mid\phi)}{\partial\phi^2}\right) = -\mathbb{E}_{X|\theta}\left(\frac{\partial^2 \log f(X\mid\phi)}{\partial\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) = I(\theta)\left|\frac{d\theta}{d\phi}\right|^2$$

Jeffrey's Noninformative Priors

- For example, consider normally distributed data with unknown mean.

- Likelihood

$$x_i \mid \theta \sim \mathcal{N}(\theta, \sigma^2)\ (known\ \sigma)$$

i.e.

$$f(x_{1:n}|\theta) \propto exp\left(-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\right),\ where\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Then:

$$\frac{\partial^2 \log f(x_{1:n} \mid \theta)}{\partial \theta^2} = -\frac{n}{\sigma^2} \Rightarrow \quad \pi(\theta) \propto 1$$

Jeffrey's Noninformative Priors

- Consider normally distributed data with unknown variance
  - Likelihood $X_i \mid \theta \sim \mathcal{N}(\mu, \theta)\ (known\ \mu)$

    i.e.

$$f(x_{1:n}|\theta) \propto \theta^{-n/2} exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2 \ \ + n(\bar{x}-\mu)^2}{2\theta}\right)$$

Then: $$\frac{\partial^2 \log f(x_{1:n}|\theta)}{\partial\theta^2} = \frac{n}{2\theta^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2 \ \ + n(\bar{x}-\mu)^2}{\theta^3} \Rightarrow$$

$$I(\theta) = -\mathbb{E}_{X|\theta}\left(\frac{n}{2\theta^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2 \ \ + n(\bar{x}-\mu)^2}{\theta^3}\right) = -\frac{n}{2\theta^2} + \mathbb{E}_{X|\theta}\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\theta^3}\right)$$

$$= -\frac{n}{2\theta^2} + \frac{n}{\theta^2} = \frac{n}{2\theta^2}$$

  - Jeffrey's prior $\pi(\theta=\sigma^2) \propto \frac{1}{\theta} = \frac{1}{\sigma^2}$ (favors small variance)
  - Note that $\pi(\phi = \log\theta) \propto \frac{1}{\theta}\left|\frac{d\theta}{d\phi}\right| = \frac{1}{\theta}\theta = 1$

Jeffrey's Noninformative Priors

- Consider data following a binomial distribution (mean $n\theta$)
  - Likelihood

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Then:

$$\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \Rightarrow I(\theta) = -\mathbb{E}_{X|\theta}\left(-\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\right) = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}$$

  - The Jeffrey's prior is:

$$\boxed{\pi(\theta) \propto [\theta(1-\theta)]^{-1/2} = \mathcal{B}eta\left(\theta; \frac{1}{2}, \frac{1}{2}\right)}$$

Beta Distribution:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$\theta \in [0,1]$$

$$\mathrm{E}(\theta) = \frac{\alpha}{\alpha+\beta}$$
$$\mathrm{var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$
$$\mathrm{mode}(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$$

- For a multinoulli random variable with K states, one can show that the Jeffreys' prior is:

$$\boxed{\pi(\theta) = \mathcal{D}ir\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)}$$

- Note that this is not any of the expected answers:

$$\pi(\theta) = \mathcal{D}ir\left(\frac{1}{K}, \ldots, \frac{1}{K}\right) \text{ or } \pi(\theta) = \mathcal{D}ir(1, \ldots, 1)$$

Pros and Cons of Jeffrey's Priors

- It can lead to incoherencies; e.g. the Jeffrey's prior for Gaussian data and $\theta = (\mu, \sigma)$ unknown is $\pi(\theta) \propto \sigma^{-2}$. Indeed using: $\ln f(x|\theta) = \ln \frac{1}{(2\pi)^{1/2}} - \ln\sigma - \frac{1}{2\sigma^2}(x-\mu)^2$

$$I(\theta) = \mathbb{E}_{X|\theta}\begin{bmatrix} \frac{1}{\sigma^2} & \frac{2(x-\mu)}{\sigma^3} \\ \frac{2(x-\mu)}{\sigma^3} & \frac{3(\mu-x)^2}{\sigma^4} - \frac{1}{\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \Rightarrow \pi(\theta) \propto \frac{1}{\sigma^2}$$

- However if these parameters are assumed a priori independent (using the results derived earlier) then $\pi(\theta) \propto \sigma^{-1}$.

- Automated procedure that however cannot incorporate any "physical" information.

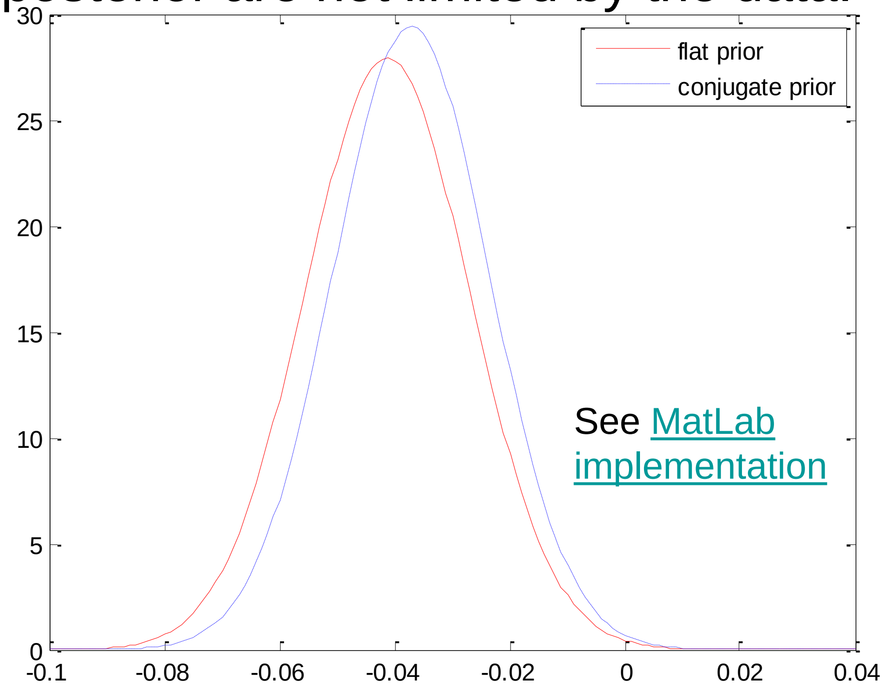
- It does NOT satisfy the likelihood principle. The Fisher information can differ for two experiments providing proportional likelihoods. For an example consider the Binomial and Negative Binomial distributions.

C. P. Robert, *The Bayesian Choice*, *Springer, 2nd edition*, chapter 3 (full text available)

Lack of Robustness of the Normal Prior

- Comparison of two posterior distributions corresponding to the flat prior (plain) and a conjugate prior (dotted) $\mathcal{N}\left(0, 0.1\bar{\sigma}^2\right)$ (where the variance $\bar{\sigma}^2$ refers here to the empirical variance of the sample). We use the data normaldata. This shows the lack of robustness of the normal prior.

- When the hyperparameters in the prior vary, both the range and location of the posterior are not limited by the data.



See MatLab implementation

J.-M. Marin & C. P. Robert, *The Bayesian Core, Springer, 2nd edition,* chapter 2 (full text available)

Robust Priors: Priors with Heavy Tails

- In many cases, we are not very confident in our prior, so we want to make sure it does not have an undue influence on the result.
- This can be done by using *robust priors, which typically have heavy tails, which avoids forcing things to be too close to the prior mean*.
- As an example, consider $x \sim \mathcal{N}(\theta, 1)$. We observe that $x = 5$ and we want to estimate θ. The MLE is $\bar{\theta} = 5$, which seems reasonable. The posterior mean under a uniform prior is also $\mathbb{E}[\theta \,|\, x = 5] = 5$.
- Suppose we know that the prior median is 0, and the prior quantiles are at -1 and 1, so $p(\theta \leq -1) = p(-1 < \theta \leq 0) = p(0 < \theta \leq 1) = p(1 < \theta) = 0.25$. Let us also assume the prior is smooth and unimodal.
- Using the prior $\mathcal{N}(\theta|0, 2.19^2)$ satisfies these prior constraints. But in this case the posterior mean is 3.43, which is not very satisfactory.
- Use Cauchy prior $\mathcal{T}(\theta|0,1,1)$. This also satisfies the prior constraints of our example. But this time we find that the posterior mean is about 4.6, which seems much more reasonable.

robustPriorDemo
from Kevin Murphys' PMTK

Hierarchical Bayesian Models

- It often helps to decompose prior knowledge into several levels particularly when the available data is hierarchical.

- The **hierarchical Bayes method** is a powerful tool for expressing rich statistical models that more fully reflect a given problem than a simpler model could.

- Often the prior on θ depends in turn on other parameters ϕ that are not mentioned in the likelihood. So, the prior $\pi(\theta)$ must be replaced by a prior $\pi(\theta|\phi)$, and a prior $\pi(\phi)$ on the newly introduced parameters ϕ is required, resulting in a posterior probability $\pi(\theta,\phi|\boldsymbol{x})$.

$$\pi(\theta,\phi \mid \mathbf{x}) \sim \underbrace{\pi(\mathbf{x} \mid \theta,\phi)}_{\pi(\mathbf{x}|\theta)} \pi(\theta,\phi) \sim \pi(\mathbf{x} \mid \theta) \underbrace{\pi(\theta \mid \phi)\pi(\phi)}_{\pi(\theta,\phi)}$$

- This is the simplest example of a *hierarchical Bayes model*.

- The process may be repeated, e.g, ϕ may depend on parameters ψ, which will require their own prior. **Eventually the process must terminate**, with priors that do not depend on any other parameters.

Hierarchical Bayesian Models

- Consider m –level hierarchical Bayesian model

$$\pi(\theta) = \int_{\Theta_1 \times \Theta_1 \times \ldots \times \Theta_m} \pi(\theta \mid \theta_1)\pi(\theta_1 \mid \theta_2)\ldots\pi(\theta_{m-1} \mid \theta_m)\pi(\theta_m)d\theta_1\ldots d\theta_m$$

- Two level hierarchical modeling gives:
  - ✓ Full posterior: $\pi(\theta, \theta_1 \mid \boldsymbol{x}) \sim \underbrace{\pi(\boldsymbol{x} \mid \theta)\pi(\theta \mid \theta_1)}_{\pi(\theta|\theta_1, \boldsymbol{x})}\pi(\theta_1)$
  - ✓ Conditional posterior: $\pi(\theta \mid \theta_1, \boldsymbol{x}) \sim \pi(\boldsymbol{x} \mid \theta)\pi(\theta \mid \theta_1)$
  - ✓ Marginal Posterior: $\pi(\theta \mid \boldsymbol{x}) = \int \pi(\theta, \theta_1 \mid \boldsymbol{x})d\theta_1$

Hierarchical Bayes

- A key requirement for computing the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is the specification of a prior $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ are the hyper-parameters.

- What if we don't know how to set $\boldsymbol{\eta}$?

- In some cases, we can use uninformative priors as discussed earlier.

- A more Bayesian approach is to put a prior on our priors! *In terms of graphical models (showing explicitly dependence relations)*, we can represent the situation as follows:

$$\boldsymbol{\eta} \rightarrow \boldsymbol{\theta} \rightarrow \mathcal{D}$$

- This is an example of a hierarchical Bayesian model, also called a multi-level model, since there are multiple levels of unknown quantities.

Hierarchical Bayes: Modeling Cancer Rates

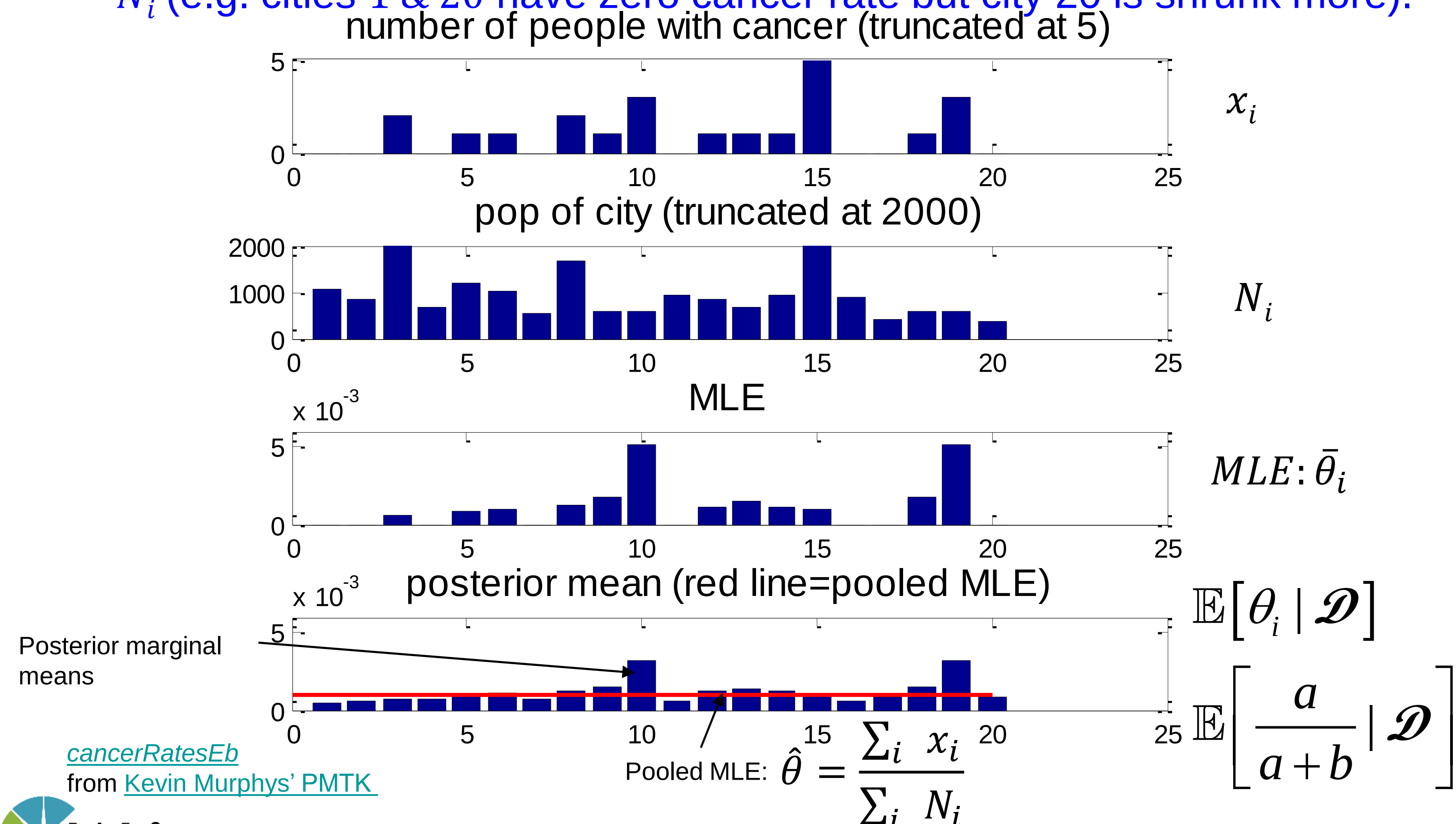
- Consider the problem of predicting cancer rates in various cities.
- We measure the people in various cities, N_i, and the people who died of cancer in these cities, x_i. We assume $x_i \sim \mathcal{Bin}(N_i, \boldsymbol{\theta}_i)$ and we estimate the cancer rates $\boldsymbol{\theta}_i$.
- We can estimate them all separately, but this will suffer from the sparse data problem (underestimation of the rate of cancer due to small N_i).
- We can assume all the $\boldsymbol{\theta}_i$ are the same (*parameter tying*). But the assumption that all the cities have the same rate is a rather strong one.
- As a compromise we assume that the $\boldsymbol{\theta}_i$ are similar, but that there may be city-specific variations. This can be modeled by assuming $\boldsymbol{\theta}_i \sim \mathcal{Beta}(a, b)$. The full joint distribution can be written as

$$p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{\eta}) \prod_{i=1}^{N} \mathcal{Bin}(x_i \mid N_i, \boldsymbol{\theta}_i) \mathcal{Beta}(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}), \boldsymbol{\eta} = (a, b)$$

- By treating $\boldsymbol{\eta}$ as an unknown (hidden variable), we *allow the data-poor cities to borrow statistical strength from data-rich ones.*

Hierarchical Bayes: Modeling Cancer Rates

- Compute $p(\eta,\theta|\mathcal{D})$, then the marginal $p(\theta|\mathcal{D})$. The posterior mean is shrunk towards the pooled estimate more strongly for cities with small N_i (e.g. cities 1 & 20 have zero cancer rate but city 20 is shrunk more).

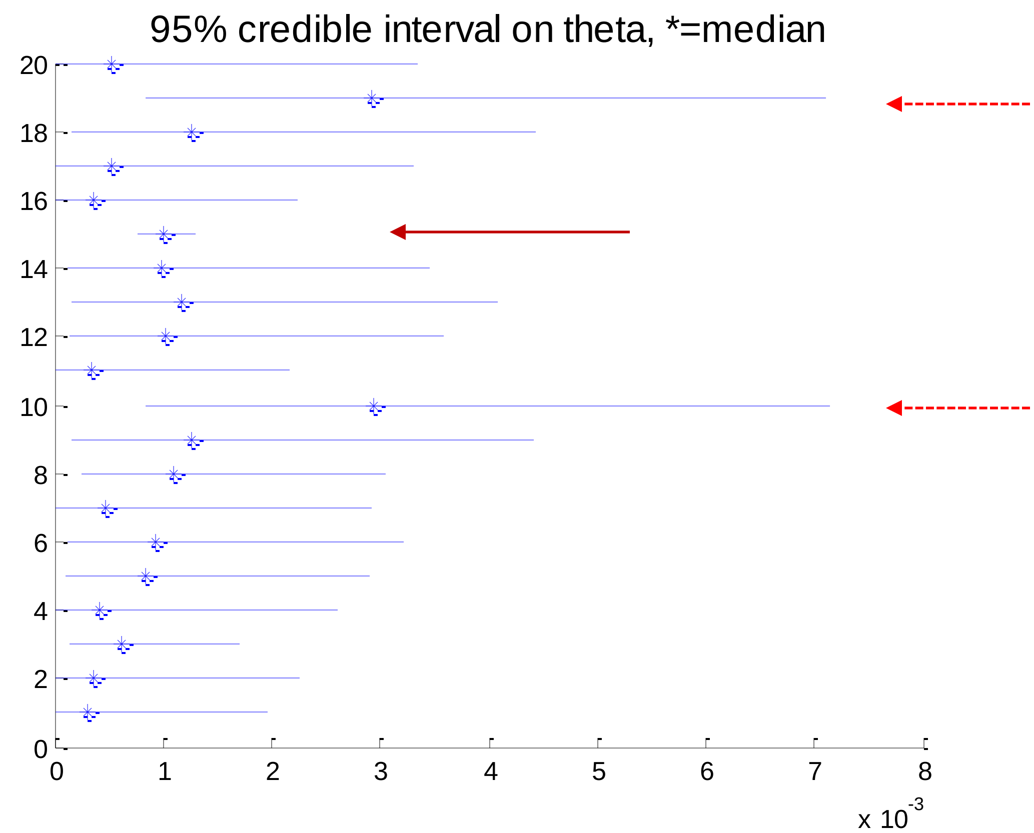


cancerRatesEb from Kevin Murphys' PMTK

Hierarchical Bayes: Modeling Cancer Rates

- 95% posterior credible intervals for $\boldsymbol{\theta}_i$.
- City 15, which has a very large population, has small posterior uncertainty. It has the largest impact on the posterior of $\boldsymbol{\eta}$ which in turn impacts the estimate of the cancer rates for other cities.
- Cities 10 and 19, which have the highest MLE, also have the highest posterior uncertainty, reflecting the fact that such a high estimate is in conflict with the prior (which is estimated from all the other cities).

[cancerRatesEb from Kevin Murphys' PMTK]()



Empirical Bayes – Evidence Approximation

- In hierarchical Bayesian models, we need to compute the posterior on multiple levels of latent variables. For example, in a two-level model,

$$p(\boldsymbol{\eta},\boldsymbol{\theta}\,|\,\mathcal{D}) \propto p(\mathcal{D}\,|\,\boldsymbol{\theta})\,p(\boldsymbol{\theta}\,|\,\boldsymbol{\eta})\,p(\boldsymbol{\eta})$$

- In some cases, we can analytically marginalize out θ; this leaves is with the simpler problem of just computing $p(\boldsymbol{\eta}\,|\mathcal{D})$.

- As a computational shortcut, we can *approximate the posterior on the hyper-parameters with a point-estimate,*

$$p(\boldsymbol{\eta}|\mathcal{D}) \approx \delta_{\overline{\boldsymbol{\eta}}}(\boldsymbol{\eta}),\ \overline{\boldsymbol{\eta}} = \operatorname{argmax} p(\boldsymbol{\eta}|\mathcal{D}) = \operatorname{argmax}\left[\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}\right]$$

- Since $\boldsymbol{\eta}$ is typically much smaller than $\boldsymbol{\theta}$ in dimensionality, it is less prone to overfitting, so we can safely use a uniform prior on $\boldsymbol{\eta}$.

- The quantity inside the brackets is *the marginal likelihood, often called the evidence.* The approach is called *empirical Bayes (EB) or type-II maximum likelihood* or the *evidence procedure*.

Empirical Bayes

- *Empirical Bayes violates the principle that the prior should be chosen independently of the data.*

- We can just view it as a cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model $\boldsymbol{\theta} \to \mathcal{D}$.

- We can construct a hierarchy in which the more integrals one performs, the "more Bayesian" one becomes:

Method	Definition
Maximum likelihood	$\hat{\theta} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\vert\theta)$
MAP estimation	$\hat{\theta} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\vert\theta)p(\theta\vert\eta)$
ML-II (Empirical Bayes)	$\hat{\eta} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\vert\theta)p(\theta\vert\eta)d\theta = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\vert\eta)$
MAP-II	$\hat{\eta} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\vert\theta)p(\theta\vert\eta)p(\eta)d\theta = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\vert\eta)p(\eta)$
Full Bayes	$p(\theta, \eta\vert\mathcal{D}) \propto p(\mathcal{D}\vert\theta)p(\theta\vert\eta)p(\eta)$

Empirical Bayes

- Let us return to the cancer rates model. We can analytically integrate out $\boldsymbol{\theta}_i$, and write down the marginal likelihood directly, as follows:

$$p(\mathcal{D}\,|\,a,b)=\prod_i \int \mathcal{B}in(x_i\,|\,\boldsymbol{N}_i,\boldsymbol{\theta}_i)\,\mathcal{B}eta(\boldsymbol{\theta}_i\,|\,a,b)\,d\boldsymbol{\theta}_i=\prod_i \binom{N_i}{x_i}\frac{B(a+x_i,b+N_i-x_i)}{B(a,b)}$$

- Various ways of maximizing this wrt a and b are discussed in Minka.

- Having estimated a and b, we can plug in the hyper-parameters to compute the posterior $p(\theta_i|\mathcal{D},\bar{a},\bar{b})$ in the usual way, using conjugate analysis.

- It can be shown that *the posterior mean of each $\boldsymbol{\theta}_i$ is a weighted average of its local MLE and the prior means, which depends on $\boldsymbol{\eta}=(a,b)$.*

- Since $\boldsymbol{\eta}$ is estimated using all the data, *each $\boldsymbol{\theta}_i$ is influenced by all data.*

Minka, T. (2000e). Estimating a Dirichlet distribution, Technical Report.

Empirical Bayes: Gaussian-Gaussian Model

- We now consider an example where the data is real-valued. We use *a Gaussian likelihood and a Gaussian prior.*

- Suppose we have data from multiple related groups, e.g. x_{ij} is the test score for student i in school j, $j = 1{:}D, i = 1{:}N_j$. *We want to estimate the mean score for each school, θ_j.*

- *Since N_j may be small for some schools, we regularize the problem by using a hierarchical Bayesian model, where θ_j comes from a common prior, $\mathcal{N}(\mu, \tau^2)$.*

- The joint distribution has the following form:

$$p\left(\boldsymbol{\theta}, \mathcal{D} \mid \boldsymbol{\eta}, \sigma^2\right) = \prod_{j=1}^{D}\left(\prod_{i=1}^{N_j}\mathcal{N}\left(x_{ij} \mid \theta_j, \sigma^2\right)\mathcal{N}\left(\theta_j \mid \mu, \tau^2\right)\right), \boldsymbol{\eta} = (\mu, \tau)$$

- We assume for simplicity that σ^2 is known.

Empirical Bayes: Gaussian-Gaussian Model

- We rewrite the joint distribution exploiting the fact that *N_j Gaussian measurements with values x_{ij} and variance σ^2 are equivalent to one measurement* $\bar{x}_j = \frac{1}{N_j}\sum_{i=1:N_j} x_{ij}$ *with variance* $\sigma_j^2 = \sigma^2/N_j$.
- This yields the following unnormalized posterior

$$p(\theta,\mathcal{D}|\hat{\eta},\sigma^2) = \prod_{j=1}^{D} \mathcal{N}\left(\theta_j|\hat{\mu},\hat{\tau}^2\right)\mathcal{N}\left(\bar{x}_j|\theta_j,\sigma_j^2\right)$$

where:

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D}\bar{x}_j \equiv \bar{x},\ \sigma^2 + \hat{\tau}^2 = \frac{1}{D}\sum_{j=1}^{D}\left(\bar{x}_j - \bar{x}\right)^2$$

Computed with the Evidence approximation

- From this, closing the square on θ_j, it follows that the posteriors are:

$$p\left(\theta_j|D,\hat{\mu},\hat{\tau}^2\right) = \mathcal{N}\left(\theta_j|\hat{B}_j\hat{\mu} + \left(1-\hat{B}_j\right)\bar{x}_j, \left(1-\hat{B}_j\right)\sigma_j^2\right),$$

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D}\bar{x}_j \equiv \bar{x},\ \sigma^2 + \hat{\tau}^2 = \frac{1}{D}\sum_{j=1}^{D}\left(\bar{x}_j - \bar{x}\right)^2 \quad ,\hat{B}_j = \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2}$$

Empirical Bayes: Gaussian-Gaussian Model

- Note that for constant σ_j^2 we can compute the evidence:

$$\int \mathcal{N}\left(\theta_j|\mu,\tau^2\right)\mathcal{N}\left(\bar{x}_j|\theta_j,\sigma^2\right)d\theta_j = \mathcal{N}\left(\bar{x}_j|\mu,\sigma^2+\tau^2\right) \Rightarrow$$

$$p(\mathcal{D}|\mu,\tau^2,\sigma^2) = \prod_{\mathrm{j}=1}^{D}\mathcal{N}\left(\bar{x}_j|\mu,\sigma^2+\tau^2\right)$$

- We can now derive the previously shown estimates using MLE:

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D}\bar{x}_j \equiv \bar{x}\,,\ \sigma^2+\hat{\tau}^2 = \frac{1}{D}\sum_{j=1}^{D}\left(\bar{x}_j-\bar{x}\right)^2 \equiv s^2$$

- In general use $\hat{\tau}^2 = max\{0, s^2-\sigma^2\}$

- For non-constant σ_j^2, you need to use Expectation-Maximization to derive the empirical Bayes (EB) estimate. Full Bayesian inference is also possible.

James Stein Estimator

$$p\left(\theta_j | D, \hat{\mu}, \bar{\tau}^2\right) = N\left(\theta_j | \hat{B}_j \hat{\mu} + \left(1 - \hat{B}_j\right)\bar{x}_j, \left(1 - \hat{B}_j\right)\sigma_j^2\right),$$

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D} \bar{x}_j \equiv \bar{x}\,,\ \sigma_j^2 + \hat{\tau}^2 = \frac{1}{D}\sum_{j=1}^{D}\left(\bar{x}_j - \bar{x}\right)^2 \quad , \hat{B}_j = \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2}$$

- The quantity $0 \leq \hat{B}_j \leq 1$ controls the degree of *shrinkage towards the overall mean,* $\hat{\mu}$.

- If the data is reliable for group j, then σ_j^2 will be small relative to $\hat{\tau}^2$; hence $\hat{B}_j$ will be small, and we will put more weight on $\bar{x}_j$ when we estimate θ_j. However, groups with small N_j will get regularized (shrunk towards the overall mean $\hat{\mu}$) more heavily.

- For σ_j constant across j, the posterior mean becomes (James Stein estimator):

$$\hat{\theta}_j = \hat{B}\bar{x} + \left(1 - \hat{B}\right)\bar{x}_j = \bar{x} + \left(1 - \hat{B}\right)\left(\bar{x}_j - \bar{x}\right),\ \hat{B} = \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2}$$

Predicting Baseball Scores

- This is an example of shrinkage applied to baseball batting averages.
- We observe the number of hits for $D = 18$ players during the first $T = 45$ games. Let the number of hits b_i and assume $b_j \sim \mathcal{B}(T, \theta_j)$, where θ_j is the "true" batting average for player j. The goal is to estimate the θ_j.
- The MLE is $\hat{\theta}_j = x_j$, $x_j = b_j / T$ being the empirical batting average. One can use an Empirical Bayes approach to do better.
- To apply the Gaussian shrinkage approach described above, we require that the likelihood be Gaussian, $x_j \sim \mathcal{N}(\theta_j, \sigma^2)$ for known σ^2.
- However, in this example we have a binomial likelihood. While this has the right mean, $\mathbb{E}[x_j] = \theta_j$, the variance is not constant:

$$\mathrm{var}\left[x_j\right] = \frac{1}{T^2}\mathrm{var}\left[b_j\right] = \frac{T\theta_j\left(1-\theta_j\right)}{T^2}$$

- Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *J. of the Am. Stat. Assoc. 70*(350), 311–319.

Predicting Baseball Scores

- So we apply a variance stabilizing transform to x_j to better match the Gaussian assumption.

$$y_j = f(x_j) = \sqrt{T}\,\arcsin\left(2x_j - 1\right)$$

- Now we have approximately $y_j \sim \mathcal{N}(f(\theta_j), 1) = \mathcal{N}(\mu_j, 1)$. We use Gaussian shrinkage to estimate the μ_j using

$$\hat{\mu}_j = \hat{B}\bar{y} + \left(1 - \hat{B}\right)\bar{y}_j = \bar{y} + \left(1 - \hat{B}\right)\left(\bar{y}_j - \bar{y}\right)$$

with $\sigma^2 = 1$, and we then transform back to get

$$\hat{\theta}_j = 0.5 sin\left(\frac{\hat{\mu}_j}{\sqrt{T}} + 1\right)$$

- The results are shown next.

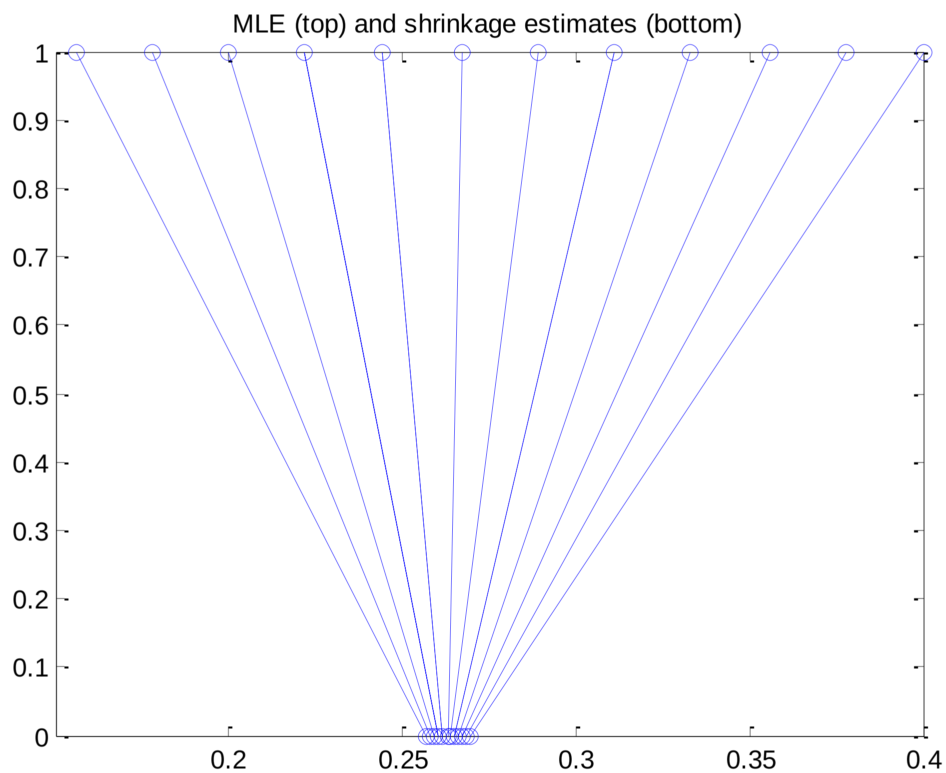
$$\textit{Consider a transform } Y = f(X) \textit{ where } \mathbb{E}[X] = \mu, \mathrm{var}[X] = \sigma^2 \textit{ s.t.}$$

$$Y = f(X) \approx f(\mu) + f'(\mu)(X - \mu) \textit{ with } \mathrm{var}[Y] = f'(\mu)^2 \sigma^2(\mu)$$

$$\textit{If } f'(\mu)^2 \sigma^2(\mu) \textit{ is independent of } \mu, \textit{ we call } f(X) \textit{ a variance stabilizing transform}$$

$$\textit{Here}: f'(\mu)^2 \sigma^2(\mu) = \left.\frac{4T}{1 - \left(2x_j - 1\right)^2}\right|_{\mu = \mathbb{E}[x_j] = \theta_j} \frac{T\theta_j\left(1 - \theta_j\right)}{T^2} = 1$$

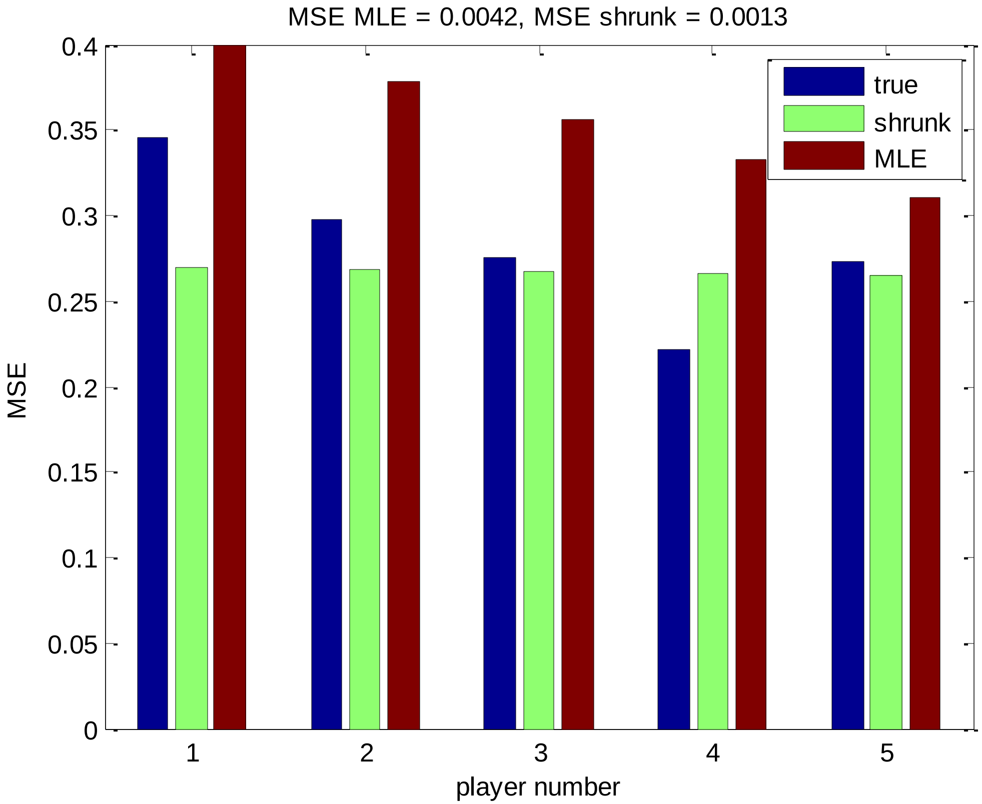
Predicting Baseball Scores



ShrinkageDemoBaseBall
from Kevin Murphys' PMTK

- We plot the MLE $\hat{\theta}_j$.
- All the estimates have shrunk towards the global mean, 0.265.

Predicting Baseball Scores



ShrinkageDemoBaseBall
from Kevin Murphys' PMTK

$$MSE = \frac{1}{N}\sum_{j=1}^{D}\left(\theta_j - \bar{\theta}_j\right)^2$$

- We plot the true value θ_j, the MLE $\hat{\theta}_j$ and the posterior mean $\bar{\theta}_j$
- The "true" values of θ_j are estimated from a large number of independent games. On average, the shrunken estimate is much closer to the true parameters than the MLE is.
- The mean squared error is over three times smaller using the $\bar{\theta}_j$ shrinkage estimates than using the MLEs $\hat{\theta}_j$